Desde la caja negra a la comprensión: fortaleciendo el vínculo modelo-fenómeno mediante inteligencia artificial explicable*

From the Black Box to Understanding: Strengthening the Model-Phenomenon Link through Explainable Artificial Intelligence

Dana Andrea García Carrillo†

Resumen

Este artículo aborda el desafío de obtener comprensión científica a partir de modelos de aprendizaje automático. Tomando como base a Sullivan (2022), argumento que los modelos opacos pueden proporcionar comprensión si se fortalece el vínculo con el fenómeno representado. Para ello, propongo integrar métodos de inteligencia artificial explicable con el fin de identificar características relevantes en el *input* y, basándose en ellas, recolectar evidencia empírica guiada por conocimiento teórico para reducir la incertidumbre.

Palabras clave: comprensión científica - incertidumbre del vínculo - modelos opacos - inteligencia artificial explicable

Abstract

This article addresses the challenge of deriving scientific understanding from machine learning models. Building on Sullivan (2022), I argue that opaque models can provide understanding if the link to the phenomenon they represent is strengthened. To this end, I propose integrating explainable artificial intelligence methods to identify relevant features in the input and, based on these, collect empirical evidence guided by theoretical knowledge to reduce uncertainty.

Keywords: scientific understanding - link uncertainty - black box models - explainable artificial intelligence

* Recibido: 11 de septiembre de 2025. Aceptado con revisiones: 1 de abril de 2026.

† Programa de Posgrado en Filosofía de la Ciencia, Universidad Nacional Autónoma de México (UNAM). Para contactar a la autora, por favor, escribir a: dana.gac7@comunidad.unam.mx.

Metatheoria 16(2)(2026): 63-74. ISSN 1853-2322. eISSN 1853-2330.

1. Introducción

Uno de los debates con creciente popularidad en la Filosofía de la Ciencia gira en torno a los modelos de aprendizaje automático (en adelante ML, del inglés *machine learning*). Estos modelos son considerados herramientas eficaces en la resolución de tareas complejas en diversos contextos, tanto científicos como sociales. Pero, pese al aparente éxito de sus resultados, también se ha cuestionado si comprendemos realmente los fenómenos que están representando. La comprensión es un logro epistémico que ocurre cuando se establecen de manera exitosa diferentes correlaciones que estén empírica y teóricamente justificadas y que permitan realizar inferencias. En el contexto del ML, es posible distinguir entre comprender los procesos internos del algoritmo y entre establecer una relación justificada entre los *outputs* del modelo y el fenómeno estudiado. Este artículo aborda esta última, la cual no es una propiedad intrínseca del modelo, ni conlleva el acceso completo a su funcionamiento interno. Más bien, se trata de una comprensión pragmática, un logro que no requiere un isomorfismo perfecto entre modelo y fenómeno, sino de representaciones *lo suficientemente verdaderas* que doten al agente de la destreza necesaria para situar el fenómeno en un espacio de posibilidades y orientar la investigación científica.

Al respecto, Sullivan (2022) propone la *opacidad de cajas negras de implementación*, la cual está dividida en tres niveles diferentes: i) de bajo nivel, en este se ocultan funciones básicas del sistema. Este por sí mismo no impide la comprensión, siempre que los datos estén relacionados adecuadamente con el fenómeno; ii) de nivel medio, la opacidad surge debido a las actualizaciones del programa durante su ejecución, pues el programador no controla directamente estos cambios. Este tampoco impide la comprensión del fenómeno siempre que la explicación se base en propiedades de nivel macro del algoritmo; iii) de alto nivel, esto ocurre cuando solo se conocen los datos de entrada y salida del algoritmo, pero todos los detalles internos permanecen ocultos para el usuario.

No obstante, en la perspectiva de Sullivan, la dificultad para alcanzar la comprensión del fenómeno va más allá de la opacidad de implementación del modelo. En cambio, la principal barrera es la *incertidumbre del vínculo*, la cual se refiere a la falta de evidencia empírica que respalde el vínculo que relaciona al modelo con el fenómeno de interés. En otras palabras, el no comprender por completo los detalles internos de la implementación de un algoritmo no es lo que impide la comprensión del fenómeno que este representa. Lo que sí la impide es la incertidumbre sobre si las relaciones identificadas por el modelo corresponden fielmente con el mundo real.

A partir de la propuesta de Sullivan, distingo entre dos formas de incertidumbre del vínculo, mismas que no deben considerarse como mutuamente excluyentes ya que en la práctica pueden afectar de manera conjunta a un modelo opaco, a saber: i) por diseño, y ii) porque el *output* no es interpretable en términos del fenómeno. El primer tipo refiere a la deficiencia estructural del modelo. Esto ocurre cuando, desde su arquitectura o datos de entrenamiento, no incorpora adecuadamente las propiedades relevantes del fenómeno real. El segundo tipo de incertidumbre se debe a que el modelo produce resultados que no pueden conectarse con el fenómeno de interés porque no sabemos cómo el resultado está representando algo de él. En particular, este tipo de incertidumbre es la de mi interés, no porque descarte la importancia del primer tipo, sino porque reconozco que aun cuando el modelo es técnicamente sólido, puede existir una barrera entre éste y el fenómeno si la evidencia empírica no es suficiente.

Por otro lado, han surgido métodos denominados *Inteligencia Artificial Explicable* (en adelante XAI), como mapas de saliencia o técnicas de atribución por capas, mismos que permiten identificar qué partes del *input* han resultado más relevantes para la producción del *output*.

En lo que sigue defiendo que los modelos opacos pueden proporcionar comprensión científica a través del vínculo que los relaciona con el fenómeno que representan. Sin embargo, cuando la incertidumbre de vínculo surge porque los *outputs* no son interpretables en términos del fenómeno real debido a la falta de evidencia, la integración de métodos de XAI puede contribuir a identificar las características más relevantes del algoritmo y con ello fortalecer el vínculo a través de la integración de evidencia empírica, la cual es guiada por el conocimiento teórico del fenómeno. El esquema argumentativo que seguiré es el siguiente. En la primera sección abordo cómo podemos obtener comprensión de los fenómenos a partir de modelos de ML, retomando la propuesta de Sullivan (2022).

En la segunda sección exploro la importancia de reducir la incertidumbre del vínculo. En la tercera sección desarrollo mi propuesta, integrar métodos de XAI con orientación teórica para fortalecer el vínculo modelo-fenómeno. La cuarta sección aborda un estudio de caso en econometría, donde se utilizan modelos de redes neuronales profundas y métodos de XAI. Finalmente, la quinta sección discute posibles objeciones a mi propuesta y les brinda una respuesta.

2. Comprensión a partir de modelos de aprendizaje automático

Los modelos de ML han cambiado a los campos científicos al facilitar el procesamiento de datos a gran escala. Sin embargo, su incorporación en la investigación científica suscita dudas sobre si pueden producir conocimiento con rigor epistémico. Un problema persistente es la opacidad de estos modelos, que dificulta desentrañar los procesos que generan sus resultados. Aunque esta característica suele percibirse como una limitación para entender los fenómenos estudiados, Sullivan (2022) argumenta que el verdadero impedimento reside en la *incertidumbre del vínculo*; esto es, a la insuficiencia de evidencia empírica que relacione los resultados del modelo con el fenómeno representado.

Sullivan argumenta que no todas las formas de opacidad tienen el mismo peso epistémico, un tipo particular de esta opacidad es la *caja negra de implementación*, la cual se descompone en tres niveles diferentes:

- Nivel bajo, alude a aquellas funciones básicas del sistema que permanecen ocultas para el usuario.
- Nivel medio, se refiere a una opacidad más dinámica: emerge cuando los parámetros del modelo se ajustan automáticamente durante la ejecución o el proceso de entrenamiento, sin intervención explícita por parte del programador.
- Nivel alto, describe situaciones en las que el usuario solo tiene acceso a los *inputs* y *outputs* del modelo, mientras que los procesos intermedios permanecen oscurecidos. Este es el único nivel que representa una barrera para la comprensión, ya que no se tiene información alguna sobre el funcionamiento interno del modelo, lo que imposibilita la justificación sobre los resultados de este.

Con esto, se establece la diferencia entre comprender cómo funciona el modelo con respecto a su implementación, y entre usar ese modelo para comprender un fenómeno de interés, pues lo que limita la comprensión no es la opacidad interna del modelo, sino la incertidumbre del vínculo. Es así como lo verdaderamente problemático no es transparentar al modelo, más bien la solidez de la relación de sus datos de salida con el mundo empírico, la cual se ve disminuida cuando la evidencia no es suficiente para respaldarla. Lo anterior debido a que debemos tener razones científicamente válidas para interpretar los *outputs* del modelo no como meras correlaciones, sino como afirmaciones con valor epistémico sobre la realidad.

En relación con lo anterior encontramos posturas como la de Durán y Pozzi (2025) quienes proponen el *Confiabilismo Computacional* como una alternativa para justificar la fiabilidad en los algoritmos de inteligencia artificial.[1] Este enfoque se aleja de los sistemas de cajas negras, y en su lugar sostiene que la fiabilidad depende de prácticas establecidas, conocimiento concerniente al diseño, desarrollo y uso de los algoritmos. El Confiabilismo Computacional acepta errores ocasionales y clasificaciones erróneas, siempre y cuando el algoritmo se mantenga generalmente fiable mediante la producción de resultados con valor científico. Esta perspectiva busca retornar a metodologías y prácticas establecidas, pero con el cuidado de integrar principios aceptados de diseño y mantenimiento algorítmico. De esta manera se aumenta la fiabilidad en estos modelos, lo cual justifica la creencia en los resultados, sin necesidad de transparentar al modelo.

[1] El Confiabilismo Computacional es una postura en la cual los resultados generados por sistemas computacionales pueden ser epistémicamente confiables, y por tanto servir como base para el conocimiento científico, si el proceso computacional que los produce es fiable, aun cuando el sistema sea en parte opaco.

Por consiguiente, exigir transparencia total en todos los modelos científicos no es solo innecesario, sino también conlleva problemas epistémicos. En la inteligencia artificial contemporánea, lo relevante no es acceder a cada proceso interno del modelo, sino establecer, a través de diferentes mediaciones, una relación confiable entre el modelo y el fenómeno que este representa.

Adicionalmente, es posible abordar la opacidad de los modelos desde diferentes perspectivas. En particular, Humphreys (2009) define a la *opacidad epistémica* como la dificultad de establecer una representación directa y significativa entre un sistema de inteligencia artificial y los fenómenos del mundo real que busca modelar. Es decir, surge a partir de la naturaleza abstracta de los algoritmos, mismos que no necesariamente se alinean a la perfección con los diversos matices del mundo. Sin embargo, como sugiere Sullivan (2022), no todos los niveles de opacidad obstaculizan la comprensión, sino que esto depende del uso del modelo y de las preguntas que se le planteen a este.

El algoritmo por sí mismo no es una explicación; estos solo explican cuando se utilizan para responder preguntas. Los modelos son capaces de proporcionar explicaciones de "cómo es posible" la ocurrencia de un fenómeno, i.e., pueden mostrar una posible causa. Sin embargo, para responder preguntas del tipo "por qué" o "cómo es realmente" un fenómeno en el mundo real, se requiere evidencia empírica adicional. Esto se inserta en un marco de discusión sobre el tipo de comprensión que pueden ofrecer esta clase de modelos. Más aún, mientras mayor sea el respaldo empírico que tenga el modelo, este podrá proporcionar una comprensión más profunda del fenómeno objetivo. Entonces, comprender no equivale a transparentar por completo cada proceso interno del modelo, sino que se trata de una práctica situada, en la que interactúan múltiples aspectos, tales como el tipo de preguntas que orientan la investigación, el contexto en el que se aplica el modelo, así como la teoría y la evidencia empírica.

La distinción entre comprender los mecanismos internos de un modelo y comprender el fenómeno que este representa se enriquece a partir del giro pragmático de Páez (2019, 2026), quien argumenta que el foco, dentro del contexto del ML, debe desplazarse desde la noción tradicional de explicación hacia una noción de comprensión pragmática. Esto implica que el agente es capaz de: i) extraer inferencias contrafácticas correctas sobre el objeto de estudio, lo que implica situar correctamente un fenómeno dentro de un espacio de posibilidades (Páez 2026, p.146). Una inferencia contrafáctica es un razonamiento teórico que evalúa qué sucedería en el fenómeno si se altera una variable específica; depende de la teoría para definir sus condiciones de fondo y las restricciones causales que rigen dicho espacio de posibilidades. ii) darle uso práctico al conocimiento, lo cual abarca diversas capacidades, tales como manipular el sistema o mejorarlo (Páez 2026, p.147).

Con esto, se refuerza la idea de que la comprensión no depende de transparentar el modelo, sino de contar con herramientas que permitan establecer relaciones significativas entre sus resultados y el fenómeno representado. Entonces, mientras que el enfoque de Sullivan exige evaluar la evidencia que respalda al modelo, el marco pragmático de Páez evalúa qué puede hacer el agente con el modelo, midiendo su destreza para navegar el espacio de posibilidades del fenómeno. Por lo tanto, ambas nociones se complementan operativamente, ya que la interacción exitosa del agente con el sistema funge como el mecanismo empírico mediante el cual se obtiene la evidencia necesaria para reducir la incertidumbre del vínculo, esto a través de, por ejemplo, elaborar una inferencia contrafáctica correcta.

Finalmente, si bien sostengo que el modelo debe establecer un vínculo con el fenómeno que representa, esta representación no debe evaluarse bajo las exigencias de un representacionalismo clásico. Por el contrario, siguiendo el marco pragmático, la representación asume un carácter funcional y no un isomorfismo perfecto. Como señala Páez (2019), apoyándose en Elgin (2017), los modelos científicos operan frecuentemente como "falsedades afortunadas". Es decir, son representaciones que, aunque divergen de la verdad literal al incorporar idealizaciones y abstracciones, resultan ser lo *suficientemente verdaderas* para aislar y ejemplificar las propiedades relevantes del fenómeno en cuestión (Elgin 2017, p. 268). Entonces, el logro epistémico de la comprensión no reside en verificar que los mecanismos internos del modelo sean una copia idéntica del fenómeno, sino en que el vínculo representacional sea lo suficientemente robusto como para permitir al agente realizar inferencias contrafácticas a partir de este.

3. De la caja negra a la incertidumbre del vínculo

Hasta ahora, he afirmado que los modelos opacos pueden facilitar la comprensión de un fenómeno, no por medio de su transparencia, sino por el vínculo que mantienen con dicho fenómeno. En otras palabras, la comprensión no depende de entender cómo funcionan los procesos internos del algoritmo, sino de tener buenas razones para creer que sus *outputs* capturaron características relevantes del fenómeno. No obstante, bajo la concepción pragmática que he adoptado, poseer estas buenas razones no consiste en un estado contemplativo de creencia en la exactitud de la representación, más bien, es la capacidad del agente para utilizar dichos datos de salida con el fin de interactuar exitosamente con el sistema y así, realizar inferencias contrafácticas correctas sobre el fenómeno empírico. Con esto surge la siguiente pregunta, ¿cómo se va a fortalecer esta relación para reducir la incertidumbre del vínculo?

Una de las principales dificultades con respecto a los modelos de ML es la *interpretabilidad* del modelo. Como señala Lipton (2018), este término es ambiguo en este contexto, pues abarca diferentes motivaciones que pueden ser conflictivas entre sí.[2] Clasifica, además, las técnicas de interpretabilidad en dos grandes tipos, cada una con sus respectivos riesgos: i) técnicas de transparencia (simulabilidad o transferencia algorítmica), ii) explicaciones post-hoc (mapas de saliencia, explicaciones, por ejemplo). Es decir, la interpretabilidad no es única ni autoevidente, sino que se trata de un concepto que debe ser matizado y contextualizado (Lipton 2018, p. 7).

En línea con lo anterior, Molnar *et al.* (2020) argumentan que los métodos de XAI ayudan a descubrir conocimiento, depurar o justificar el modelo y sus predicciones, y controlar y mejorar el modelo, aun cuando no haya una noción única de interpretabilidad. Asimismo, destacan la importancia de realizar evaluaciones centradas en el ser humano, tanto expertos como personas comunes, para determinar la calidad de la explicación y la capacidad de los diferentes usuarios para entender el modelo (Molnar *et al* 2020, p. 8). Por consiguiente, el vínculo entre el fenómeno y modelo puede fortalecerse mediante diversos aspectos, como la validación empírica y la alineación de los objetivos del modelo con las necesidades del mundo real.

Al respecto, Sullivan opina que los métodos de XAI, son una herramienta útil para los modeladores, pues permiten indagar en los aspectos relevantes para el modelo. Estos son valiosos para identificar casos en los que el modelo se enfoca en aspectos irrelevantes en lugar de los que en verdad son de interés; por ejemplo, los mapas de saliencia pueden ser utilizados en clasificadores de imágenes para resaltar las áreas de una imagen que el modelo consideró más relevantes (Sullivan 2022, p. 122). Además, diferentes tipos de pruebas de saliencia son suficientes para satisfacer nuestra necesidad de conocer los detalles de alto nivel respecto al funcionamiento del modelo, con lo cual se está más cerca de la comprensión. No obstante, pese a su utilidad, los métodos de XAI consisten en aproximaciones generales, con lo cual aún permanecen opacos diversos aspectos de nivel medio y bajo.

La pregunta sobre cómo fortalecer el vínculo entre modelo y fenómeno desplaza la atención desde la estructura interna del modelo hacia las condiciones externas que justifican su uso como herramienta epistémica. Esto nos incita a repensar los criterios de validación y la relación entre teoría, evidencia y modelos computacionales. La exigencia de transparencia total como condición necesaria para la comprensión de los modelos de ML, especialmente en aquellos de redes neuronales profundas, impone un ideal epistémico que rara vez se alcanza y que, en diversos contextos científicos, resulta innecesario. En consecuencia, pretender que la comprensión únicamente es posible cuando se conoce en detalle la arquitectura interna del modelo, como sus parámetros y pesos, conduce a la desconfianza en modelos que, aunque opacos, han demostrado ser eficaces y útiles. Así, en lugar de mantener esta exigencia rígida, es más beneficioso adoptar un enfoque situado y pragmático, en el que se construyan prácticas que combinen conocimiento teórico con diferentes métodos para la obtención de evidencia empírica. Este problema no es solamente técnico, sino que nos lleva a reflexionar sobre qué significa comprender en un contexto que no depende únicamente del humano, sino también de aspectos tecnológicamente mediados.

Establecer un vínculo epistémicamente justificado entre los *outputs* del modelo y el fenómeno que representa no depende exclusivamente del acceso estructural al modelo, sino de prácticas que sitúan al modelo en un marco de investigación científica, donde:

[2] Estas motivaciones abarcan aspectos como: confianza, causalidad, transferencia, informatividad y equidad, y ética.

- El conocimiento teórico proporciona un contexto interpretativo,
- Los métodos empíricos, tales como la validación cruzada externa, ofrecen respaldo observacional,
- Las herramientas post-hoc o de explicabilidad, pueden facilitar la identificación de patrones relevantes, variables influyentes o posibles errores.

Así, la comprensión no es una propiedad intrínseca del modelo, sino un logro distribuido y contextual que depende del uso reflexivo del sistema en un entorno epistémico.

4. Incertidumbre del segundo tipo y la orientación de acciones epistémicas

Bajo este enfoque se abre una posibilidad metodológica y epistémica relevante: integrar métodos de explicabilidad en los modelos de ML, no con la intención de disipar por completo su opacidad, sino con el propósito de reducir la incertidumbre del vínculo. Estas herramientas, entre ellas los mapas de saliencia o los métodos de atribución por capas, funcionan como medios para fortalecer la relación empírica entre el modelo y el fenómeno representado, en lugar de operar como dispositivos de transparencia. El foco está puesto, en particular, en una forma específica de incertidumbre, a saber, aquella que emerge cuando disponemos de información limitada o nula sobre cómo los *outputs* del modelo se relacionan con el fenómeno real.

Una caracterización general de la incertidumbre del vínculo es la ausencia de justificación suficiente para afirmar que las salidas del modelo guardan una relación confiable y significativa con lo empírico. Este tipo de incertidumbre, más la complejidad interna del algoritmo, es lo que obstaculiza la posibilidad de comprensión. En ese marco, elaboro la siguiente distinción: por un lado, hay casos en los que la incertidumbre proviene de deficiencias en el diseño o entrenamiento del modelo, esta es denominada del primer tipo. Esto puede ocurrir cuando, por ejemplo, el *dataset* de entrenamiento está sesgado. Por otro lado, a pesar de que el modelo parece funcionar de manera adecuada desde un punto de vista operativo, sus resultados carecen de interpretación científica clara, a esta la denomino del segundo tipo.

Por ejemplo, en un caso de diagnóstico médico asistido por un modelo de nuestro interés. Supongamos que el algoritmo es capaz de clasificar imágenes de ultrasonidos con alta precisión. Pero, si los patrones detectados por el modelo no corresponden con datos clínicos conocidos, o si no se dispone de estudios que validen sus resultados, entonces el *output* será opaco pese a su fiabilidad en términos operativos. Así, el problema no radica en que el algoritmo no funcione, sino en que no se puede establecer qué aspecto del fenómeno está siendo representado en el modelo.

Es entonces que la incertidumbre del vínculo que me interesa es la que clasifico como del segundo tipo, pues esta no sólo limita la comprensión, sino que obstaculiza el uso del modelo como herramienta epistémica, i.e., no permite generar, organizar o justificar conocimiento sobre un fenómeno. Desde esta perspectiva, ¿cómo ayuda la XAI para facilitar la comprensión a partir de modelos opacos cuando la incertidumbre del vínculo es de este tipo?

4.1. XAI como herramienta de reducción de la incertidumbre del vínculo

Los métodos de XAI consisten en aproximaciones generales, por ende, usarlos en modelos de ML provoca que estos permanezcan opacos en diversos aspectos de niveles medio y bajo. No obstante, mi principal preocupación no es la opacidad de los modelos, sino la manera en que estos se relacionan con el fenómeno que representan. Debido a esto, cuando la incertidumbre del vínculo surge porque los *outputs* no son interpretables en términos del fenómeno real debido a la falta de evidencia empírica, los métodos de XAI pueden integrarse con el fin de identificar las características más relevantes del algoritmo y con ello, fortalecer el vínculo. Esto mediante la recolección selectiva de evidencia empírica, la cual además es guiada por conocimiento teórico.

De manera general, la principal función de los métodos de XAI es identificar qué partes del *input* han sido consideradas con mayor relevancia en la generación del *output* del modelo, pero no exhiben de manera exhaustiva los detalles del proceso interno. Entre la diversidad de estos métodos destacan los

mapas de saliencia y las técnicas de atribución de características. Estos métodos ayudan a entender el funcionamiento general del modelo, lo cual es un paso inicial para la comprensión científica. Más aún, esta información es crucial en contextos donde el *output* no es totalmente interpretable.

Los métodos de XAI no eliminan la opacidad de estos modelos, pero desempeñan un papel crucial como mediaciones epistémicas que permiten localizar las características del *input* con mayor relevancia en la generación del *output*. Esto no equivale a abrir la caja negra, sino a crear formas de acceso que orienten el análisis hacia aspectos del modelo que reflejan rasgos potencialmente significativos del fenómeno objetivo. De esta manera, aunque no proporcionen una explicación exhaustiva del funcionamiento interno del algoritmo, estas herramientas facilitan la identificación de patrones relevantes y guían la construcción de estrategias empíricas orientadas a la validación. Esto es especialmente valioso cuando la incertidumbre del vínculo no proviene de errores en el diseño o en el entrenamiento, sino de la dificultad para interpretar los resultados del modelo. En estos casos, métodos como los mapas de saliencia permiten formular hipótesis más precisas en relación con el fenómeno. Por tanto, el aporte de las XAI no radica en ofrecer explicaciones definitivas, sino en guiar el proceso de investigación donde la evidencia empírica y el conocimiento teórico se integran para reducir dicha incertidumbre.

Una vez identificadas las zonas de interés, surge una nueva pregunta metodológica: ¿qué tipo de estrategia empírica puede adoptarse para fortalecer el vínculo alrededor de estos elementos destacados? Y, estrechamente ligado a esto, ¿cómo puede evitarse el riesgo de incurrir en *overfitting explicativo* o un sesgo de confirmación que distorsione la evaluación del modelo?[3] Estas cuestiones son desarrolladas en la siguiente sección, donde se exploran los límites y alcances de una estrategia empírica guiada por la teoría.

4.2. XAI y la orientación de acciones epistémicas

Implementar métodos de XAI no tiene como fin transparentar el algoritmo, sino reducir la incertidumbre del vínculo del segundo tipo al señalar qué zonas del fenómeno podrían estar representadas en el modelo. Además, la utilidad de estas herramientas no radica únicamente en este diagnóstico, sino en su potencial para orientar nuevas acciones epistémicas. Para que esta orientación sea fructífera, es indispensable adoptar una concepción pragmática de la comprensión.

Una vez identificadas las zonas de interés en el modelo, se abre la posibilidad de fortalecer el vínculo mediante una estrategia empírica dirigida. Esta consiste en diseñar maneras de recolección de datos, ya sea a través de experimentos o procesos estadísticos, específicamente orientados en las variables destacadas por la XAI, con el objetivo de determinar si los *outputs* efectivamente representan propiedades reales del fenómeno modelado. Su importancia radica no en qué transparentes son, sino en qué tan útilmente guían al modelo.

Algunas autoras (véase Burrell 2016, Rudin 2019) han señalado diversos riesgos al momento de aplicar métodos de XAI. Uno de ellos es el *overfitting explicativo*, es decir, una sobre interpretación de los patrones identificados, lo que lleva al usuario a entenderlos como si estos fueran representaciones reales del fenómeno. Otro riesgo son los sesgos de confirmación al seleccionar evidencia que refuerce las inferencias del modelo, pese a que estas puedan ser erróneas. Con el fin de evitar esto, las estrategias empíricas deben estar guiadas por conocimiento teórico. Así, el papel que la teoría juega es doble, mientras por un lado provee criterios para decidir si las variables destacadas por el método de XAI son plausibles en el contexto científico que le corresponde; por otro actúa como un marco regulativo para evaluar si la interpretación propuesta es coherente con el cuerpo de conocimiento existente.

Este punto ha sido enfatizado por Rudin (2019) quien argumenta que la interpretabilidad es una noción específica del dominio y que esta requiere de conocimiento estructural del mismo dominio, como causalidad, restricciones físicas o estructurales. Es decir, la teoría provee los criterios de plausibilidad de la interpretación. Desde esta perspectiva, la teoría no es un complemento para el modelo, sino que actúa

[3] Esto ocurre cuando una explicación generada por un modelo de aprendizaje automático ajusta excesivamente los datos o el comportamiento del modelo en un caso particular, pero no brinda ayuda con respecto al funcionamiento global del sistema. Es decir, la explicación es tan ajustada a ciertos datos locales que pierde su capacidad de generación.

como el principio normativo que garantiza la coherencia y legitimidad del vínculo entre modelo y fenómeno.

No obstante, colocar a la teoría como el principio normativo desencadena una legítima preocupación, el riesgo de un conservadurismo teórico que anule el potencial de descubrimiento científico mediado por modelos de ML. Ante esto, es crucial enfatizar que el papel normativo de la teoría opera en el plano metodológico del diseño de estrategias empíricas, no como una restricción rigurosa sobre los resultados. Cuando la XAI identifica una variable anómala respecto a la teoría vigente, esta debe tratarse como una hipótesis abductiva que se generó computacionalmente. En primera instancia, el conocimiento de la teoría postula la irrelevancia de dicha hipótesis; sin embargo, si al ejecutar la prueba empírica contrafáctica el fenómeno responde a la alteración de esa variable, lo cual contradice la expectativa teórica, entonces el modelo parece facilitar el descubrimiento científico genuino.

En este sentido, Doshi-Vélez y Kim (2017) proponen que los métodos de XAI no sólo deben ser evaluados a través de aspectos técnicos, sino también en función de su utilidad para facilitar la comprensión científica y la generación de conocimiento. Sugieren una taxonomía para realizar dicha evaluación:

- Fundamentada en la aplicación, esta conlleva realizar experimentos con seres humanos reales en un entorno real para evaluar la calidad de la explicación dentro del contexto de interés.
- Fundamentada en el ser humano, refiere a experimentos más simples con sujetos humanos que son apropiados para probar nociones más generales con respecto a la calidad de una explicación.
- Funcionalmente fundamentada, esta no requiere de experimentos humanos, sino que es más cercana a las interpretaciones técnicas.

Además, estos tres tipos de evaluación se informan mutuamente para crear los vínculos necesarios mediante la indicación de las métricas funcionales que reflejan el rendimiento en entornos del mundo real y la necesidad de la comprensión humana. Esto implica que una explicación generada por un método de XAI debe considerarse lo suficientemente confiable para guiar la investigación empírica y que dicha investigación esté normativamente orientada por hipótesis teóricas y estándares disciplinarios.

Por último, es importante distinguir si el vínculo fortalecido entre modelo y fenómeno ofrece una genuina comprensión explicativa. Al respecto, los métodos de XAI por sí mismos, no extraen ni revelan directamente los mecanismos causales subyacentes del mundo físico. Sus *outputs* exponen correlaciones matemáticas y, por consiguiente, si la herramienta de explicabilidad operara de manera aislada, el resultado sería una clasificación estadística robusta. Pero, dentro de la propuesta que defiendo, se transita de dicha clasificación hacia una *comprensión objetual*, donde el conocimiento teórico es el componente encargado de proveer la estructura causal de fondo. La XAI funciona heurísticamente señalando qué variables son de alta relevancia predictiva, pero es la teoría la que relaciona esas variables sobre los mecanismos causales del fenómeno real.

Al relacionar las características resaltadas por el algoritmo con los modelos causales de la teoría mediante inferencias contrafácticas, el agente no solo justifica la clasificación del fenómeno, sino que explica el comportamiento del sistema dentro de un espacio de posibilidades causales. Así, el vínculo modelo-fenómeno no se limita a constatar una correlación estadística dentro del algoritmo, sino que se consolida como un puente empíricamente justificado que relaciona las variables detectadas por la XAI hacia los mecanismos reales del mundo, siendo la teoría la encargada de dotar de sentido y evaluar dicha conexión.

Por tanto, la integración de métodos de XAI con estrategias empíricas guiadas por la teoría no constituye una solución técnica al problema de opacidad, sino un cambio al estatus epistémico del modelo. Cuando la transparencia es inalcanzable, como en los modelos de redes neuronales profundas, la función epistémica del modelo no puede basarse en la inspección de los mecanismos internos, sino en cómo se integra a la práctica científica de tal modo que el *output* tenga sentido y justificación en relación con el fenómeno. Bajo este marco, la comprensión no es dada por el acceso directo a las

causalidades del modelo, sino que es el resultado de la interacción de estos tres elementos: el modelo, la evidencia empírica y el conocimiento teórico de fondo.

5. Estudio de caso: redes neuronales profundas y XAI aplicadas a la econometría

La econometría ocupa un lugar central en la elaboración de políticas económicas y en el análisis cuantitativo de fenómenos macro y microeconómicos.[4] El avance de la inteligencia artificial, y en especial de los modelos de ML, ha transformado las herramientas disponibles para el modelado y la predicción dentro de esta disciplina, pues a diferencia de los modelos econométricos clásicos, que se basan en estructuras funcionales predefinidas y de supuestos estadísticos rígidos, los sistemas de inteligencia artificial pueden identificar patrones no lineales, dinámicos y altamente complejos en grandes volúmenes de datos.

En particular, la determinación del precio de la electricidad plantea desafíos desde el punto de vista económico y técnico. A diferencia de otros bienes, la electricidad es una mercancía no almacenable, sujeta a fluctuaciones abruptas por factores meteorológicos, estacionales y estructurales del mercado. Con el objetivo de capturar la dinámica temporal de los precios y ofrecer buenas predicciones se han utilizado herramientas econométricas como regresiones multivariadas. Sin embargo, como señalan Girish *et al.* (2023), así como Hai y Van Tuan (2024), estos modelos resultan insuficientes para capturar relaciones no lineales, interacciones de alta dimensionalidad o cambios abruptos en el mercado. En este contexto, los modelos de ML ofrecen una alternativa poderosa para el pronóstico de precios en mercados de electricidad. Su capacidad para procesar grandes volúmenes de datos históricos, meteorológicos y estructurales permite construir modelos predictivos más precisos y adaptativos. No obstante, usar estos modelos plantea también nuevos desafíos epistémicos. Muchos de estos son opacos en su estructura y funcionamiento interno, lo que dificulta la interpretación de sus resultados y a la vez, impone límites a su integración plena en procesos de decisión. Frente a esto, es necesario explorar en qué medida las técnicas de XAI pueden servir como puentes entre modelos altamente efectivos pero opacos y los fenómenos económicos que se busca comprender.

Pesenti y O'Sullivan (2025) proponen un estudio basado en redes neuronales profundas, cuya arquitectura permite capturar relaciones no lineales, efectos acumulativos y patrones temporales latentes en los datos del mercado. Utilizando datos históricos del mercado y variables como el uso de energías renovables, modelaron los precios con respecto al horario de cinco países diferentes. Estos modelos lograron altos niveles de precisión predictiva, pero su carácter opaco planteaba dudas sobre su interpretabilidad y, en consecuencia, sobre su utilidad para la comprensión de los fenómenos subyacentes en los mercados eléctricos. Entonces, integraron técnicas de XAI con el objetivo de revelar qué factores estaban siendo considerados por el algoritmo al momento de emitir una predicción. Por ejemplo, el método de gradiente se utilizó para calcular la importancia atribuida a cada entrada en la decisión del modelo y con ello, se observó cómo ciertas variables, tales como la demanda, contribuyen a un alza o baja del precio proyectado. Los datos que se obtuvieron a partir de la integración de herramientas de XAI son diversos, entre ellos:

- El uso de energías renovables tuvo una influencia importante, pero solo en determinadas horas del día, como la energía solar en el mediodía, mostrando que el modelo capta la estacionalidad horaria.
- El último precio disponible es la variable más importante para predecir los precios del día siguiente en los cinco mercados analizados. Es decir, los modelos muestran una preferencia por variables más recientes sobre un enfoque que utiliza el precio de ayer a la misma hora para predecir el de hoy. Sin embargo, pese a la relevancia del último precio, su impacto es moderado. Esto refleja que los modelos son complejos y utilizan

[4] La econometría es una disciplina del análisis económico cuantitativo, donde se utilizan modelos estadísticos para entender, estimar y predecir relaciones entre variables económicas, tales como el PIB, la inflación o el desempleo.

muchas de variables diferentes para realizar predicciones (Pesenti y O'Sullivan 2025, p. 10).

- El análisis proporciona información relevante para los operadores del mercado y para la formulación de políticas, lo cual mejora la transparencia y fomenta la confianza en los procesos de toma de decisiones impulsados por inteligencia artificial (Pesenti y O'Sullivan 2025, p. 11).

Los métodos de XAI utilizados por Pesenti y O'Sullivan no se limitan a funciones de auditoría técnica ni a la validación superficial de modelos predictivos. Por el contrario, cumplen con un papel epistémico relevante, a saber, reconstruir las estructuras de relación que el modelo ha inferido a partir de los datos, lo cual abre la posibilidad de evaluar en qué medida estas estructuras corresponden a propiedades reales del fenómeno estudiado. El modelo ya no solo proporciona información sobre el "qué sucederá", sino también el "por qué lo predice". Esto abre la opción de, por ejemplo, contrastar las relaciones que refleja el modelo con el conocimiento de expertos sobre cómo funcionan realmente los mercados eléctricos. Lo crucial es que estas funciones no están dadas por defecto en el modelo, sino que emergen sólo cuando se integran herramientas de XAI que actúan como mediadores entre la estructura interna del modelo y el fenómeno. En este sentido, la XAI no brinda información sobre el modelo en abstracto, sino del modo en que el modelo representa el fenómeno.

En consecuencia, el valor epistémico de la XAI reside en su capacidad para fortalecer el vínculo modelo-fenómeno. En lugar de aceptar las predicciones de cajas negras que se suponen confiables, se abre una vía hacia la reconstrucción parcial y pragmática del razonamiento del modelo.

6. Abordando posibles objeciones

Si bien he argumentado a favor de la integración de métodos de XAI para la reducción de la incertidumbre del vínculo y con ello, facilitar la comprensión científica en contextos de opacidad, esta tesis no está exenta de posibles objeciones. Algunas de estas se examinan a continuación.

Una primera objeción proviene de Adebayo *et al.* (2018), quienes proponen una metodología denominada *pruebas de cordura* para evaluar si los mapas de saliencia en verdad capturan información significativa del modelo. Sus hallazgos revelan información crítica, ya que muchos de los resultados generados por estos métodos no dependen de los parámetros del modelo entrenado. En particular, los mapas de saliencia producidos por métodos como Guided GradCAM o Guided Backprop no se ven afectados a los cambios que se realizan en el modelo base, lo cual sugiere que estos pueden no estar explicando nada del modelo en sí. Esto plantea una seria preocupación, pues si los mapas de saliencia son insensibles a la estructura del modelo, entonces no pueden considerarse una fuente fiable de información sobre las características realmente relevantes del *input* y, por ende, no son una guía válida para la reducción de incertidumbre del vínculo.

Para abordar el núcleo de esta objeción, propongo una defensa estructurada en dos niveles, siendo estos la adopción de un filtro normativo estricto y la recolección de evidencia empírica desde el marco pragmático previamente presentado.

En primer lugar, superar evaluaciones como las *pruebas de cordura* no es opcional, sino un prerrequisito metodológico normativo. Aquellos métodos de XAI que demuestren ser insensibles a la estructura y a los parámetros del modelo base carecen de justificación epistémica para referir al vínculo modelo-fenómeno. Por lo tanto, su uso debe ser descartado dentro de la investigación.

En segundo lugar, con el fin de mitigar un sesgo de confirmación sistemático con las herramientas que superen este primer filtro, es necesario contrastar la manipulación local de variables con el razonamiento contrafáctico genuino. La manipulación local se limita a alterar algorítmicamente el valor de una variable manteniendo las demás fijas para observar cambios en el *output*, asumiendo que las variables de entrada son causalmente independientes. Esto muestra cómo cambia el *output*, pero no explica por qué lo hace, ya que carece de una comprensión de las conexiones causales y lógicas subyacentes. Si la recolección de evidencia empírica se diseñara para buscar circularmente aquellos datos que corroboren estas alteraciones resaltadas por la XAI, se incurriría en el sesgo advertido.

Para evitarlo, y conforme con la concepción pragmática adoptada, el conocimiento teórico del dominio funge como la estructura que brinda el diseño de la estrategia empírica. Si la XAI destaca una variable o patrón específico, el agente debe usar la teoría para formular una inferencia sobre qué sucedería en el fenómeno real si se interviene dicha variable. Es en la recolección de evidencia donde se evita el sesgo. Si al intervenir empíricamente la variable el fenómeno no sufre alguna alteración, entonces el *output* de la XAI era un dato estadístico y procede a descartarse. Pero, si el fenómeno sí responde a la intervención, incluso si dicha no corresponde con las expectativas de la teoría, el agente no fracasa, sino que logra fortalecer el vínculo entre el modelo y el fenómeno.

Lo anterior concuerda con lo señalado por Molnar *et al.* (2020), quienes sostienen que estas técnicas tienen como propósito la depuración del modelo o la justificación de este y sus predicciones, así como ser una herramienta de control y mejora. Entendidas así, las XAI son herramientas de apoyo heurístico para los modelos de ML, y no un sustituto de validación y comprensión.

Por lo tanto, la respuesta a esta objeción no es desechar los métodos de XAI, sino integrarlos de manera crítica como parte de una estrategia que combine herramientas computacionales, conocimiento teórico y evaluación empírica.

Otra objeción por abordar es que la tesis defendida se enfoca únicamente en la reducción de la incertidumbre del vínculo del segundo tipo, es decir, de aquella que surge cuando los *outputs* del modelo no están suficientemente relacionados con el fenómeno por falta de evidencia, pero omite la incertidumbre del primer tipo.

Creel (2020) señala que los sistemas de ML pueden introducir sesgos si se basan en datos que ya están sesgados debido a su método de recopilación. A este tipo de problema lo llama "basura entra, basura sale". Aún más, Durán y Pozzi (2025) destacan que la falta de responsabilidad y la perpetuación de sesgos son ejemplos de cómo la validez epistémica de un modelo puede verse comprometida. Por consiguiente, tratar un solo tipo de incertidumbre del vínculo basada en la interpretación de los *outputs* no necesariamente soluciona todos los problemas del modelo.

En respuesta a esto, sostengo que la incertidumbre de vínculo del primer tipo es una fuente significativa de preocupación en el uso de modelos opacos, ya que compromete al modelo desde su origen. No obstante, la tesis defendida en este artículo se ha enfocado únicamente en la incertidumbre del segundo tipo, no porque se busque ignorar a la del primer tipo, sino que el propósito es reconocer que aun cuando el modelo es técnicamente sólido, puede mantenerse una barrera entre éste y el fenómeno si la evidencia empírica no es suficiente. Molnar *et al.* (2020) subrayan que la interpretabilidad no es una propiedad que resida exclusivamente en la arquitectura del modelo, sino que está intrínsecamente entrelazada con el contexto de su aplicación. Reconozco, por tanto, que focalizarse únicamente en la incertidumbre del vínculo, y en particular, en su manifestación cuando los *outputs* no son empíricamente interpretables, no equivale a resolver el problema general de la opacidad en los modelos de ML.

La tesis que defiendo mantiene su valor, pues apunta a que la comprensión no depende exclusivamente de la transparencia del modelo ni de su arquitectura inicial, sino también de las condiciones bajo las cuales se justifica el uso de los *outputs* como representaciones fiables del fenómeno. La falta de evidencia externa continúa siendo, incluso en modelos bien planteados, una barrera para la comprensión y por ello, los métodos de XAI, en paralelo con conocimiento teórico y estrategias empíricas, pueden desempeñar un papel valioso.

7. Conclusión

Los modelos opacos, pese a su falta de transparencia interna, pueden proporcionar comprensión científica siempre que exista un vínculo robusto entre sus *outputs* y el fenómeno que representan. Bajo el enfoque pragmático adoptado, la comprensión no se deriva del acceso total a los mecanismos internos del modelo, ni de contemplar de manera pasiva una representación fiel, sino de la destreza del agente para utilizar dicho vínculo con el fin de interactuar exitosamente con el sistema y situarlo en un espacio de posibilidades.

En particular, argumento que cuando la incertidumbre del vínculo es del segundo tipo, i.e., cuando surge a partir de la falta de interpretabilidad de los *outputs*, los métodos de inteligencia artificial explicable (XAI) pueden desempeñar un papel crucial, ya que estas herramientas permiten identificar qué rasgos del *input* son más relevantes para el modelo. Esto, a su vez, guía la recolección de evidencia empírica, la cual debe estar orientada por conocimiento teórico mediante la formulación de inferencias contrafácticas. Al proceder de esta manera, el modelo fortalece su papel como herramienta epistémica, reconociendo los límites normativos de las técnicas actuales de XAI y evitando caer en la idealización de estos sistemas.

Sin embargo, quedan preguntas importantes que deben abordarse en futuras investigaciones. Tales como: ¿cómo reducir la incertidumbre de tipo uno, causada cuando el modelo no incluye características significativas del fenómeno que pretende modelar? Resolver esta cuestión es importante no solo para clarificar el estatus epistémico de los modelos, sino para mejorar su contribución a la explicación y comprensión científica.

Bibliografía

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. y B. Kim (2018), "Sanity Checks for Saliency Maps", *Advances in Neural Information Processing Systems* 31: 1-11.

Burrell, J. (2016), "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms", *Big data & society* 3(1): 2053951715622512.

Creel, K. A. (2020), "Transparency in Complex Computational Systems", *Philosophy of Science* 87(4): 568-589.

Doshi-Velez, F. y B. Kim (2017), "Towards a Rigorous Science of Interpretable Machine Learning", *arXiv preprint arXiv*: 1702.08608.

Durán, J. M. y G. Pozzi (2025), "Trust and Trustworthiness in AI", *Philosophy & Technology* 38(1): 1-31.

Elgin, C. Z. (2017), *True Enough*, Cambridge, MA: MIT Press.

Girish, G. P., Bhagat, R., Preeti, S. H. y S. Singh (2023), "AI Models for Spot Electricity Price Forecasting —A Review", en Vasant, P. *et al*. (eds.), *Intelligent Computing and Optimization*, ICO 2023, Lecture Notes in Networks and Systems, Vol. 852, Cham: Springer, pp. 97-103.

Hai, D. H. y P. Van Tuan (2024), "AI and Econometric Modeling: Deep Reinforcement Learning in Predictive Modeling", en Kreinovich, V., Yamaka, W. y S. Leurcharusmee (eds.), *Applications of Optimal Transport to Economics and Related Topics. Studies in Systems, Decision and Control*, Vol. 556, Cham: Springer, pp. 53-60.

Humphreys, P. (2009), "The Philosophical Novelty of Computer Simulation Methods", *Synthese* 169: 615-626.

Lipton, Z. C. (2018), "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery", *Queue* 16(3): 31-57.

Molnar, C., Casalicchio, G. y B. Bischl (2020), "Interpretable Machine Learning–A Brief History, State-of-the-Art and Challenges", *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Cham: Springer, pp. 417-431.

Paez, A. (2025), "Axe the X in XAI: A Plea for Understandable AI", por publicarse en Durán, J. M. y G. Pozzi (eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*.

Páez, A. (2019), "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)", *Minds & Machines* 29: 441-459.

Pesenti, A. y A. O'Sullivan (2025), "Explaining Deep Neural Network Models for Electricity Price Forecasting with XAI", *Energy and AI* 21: 100532.

Rudin, C. (2019), "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead", *Nature Machine Intelligence* 1(5): 206-215.

Sullivan, E. (2022). "Understanding from Machine Learning Models", *British Journal for the Philosophy of Science* 73(1): 109-133.